The American Journal of Surgery 217 (2019) 306–313

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

Assessing gender bias in qualitative evaluations of surgical residents

Katherine M. Gerull [a, 1], Maren Loe [a, 1], Kristen Seiler [a], Jared McAllister [a], Arghavan Salles [b, *, 2]

[a] Washington University School of Medicine, St. Louis, MO, United States
[b] Stanford University, Stanford, CA, United States

ARTICLE INFO

Article history:
Received 15 May 2018
Received in revised form
2 September 2018
Accepted 28 September 2018

Keywords:
Gender bias
Performance review
Residency
Evaluation
Qualitative

ABSTRACT

Background: There are notable disparities in the training, recruitment, promotion, and evaluation of men and women in surgery. The qualitative assessment of surgical residents may be implicitly gender biased.
Methods: We used inductive analysis to identify themes in written evaluations of residents. We also performed a content analysis of words fitting previously defined communal, grindstone, ability, and standout categories.
Results: Differences in themes that emerged from evaluations of male and female residents were notable regarding overall performance, references to the future, professional competency, job domains, disposition and humanism, and overall tone of evaluations. Comments about men were more positive than those about women, and evaluations of men included more standout words.
Conclusions: The more positive evaluations of men may handicap women if they are seen as less likely to perform well based on these evaluations. These differences suggest that implicit bias may play a role in the qualitative evaluation of surgical residents.

Introduction

Written evaluations of performance are important in medical training, much as they are in other fields. For medical students, performance evaluations are used in the Medical Student Performance Evaluation which becomes part of their residency application. For residents, written evaluations are designed to give guidance on how performance can be improved. These data are often also used to assess whether residents are meeting their ACGME milestones and may impact residents' standing in their programs.

Within medical training, gender bias has been identified in written performance evaluations.[1–4] At the medical student training level, one group analyzed differences in the latent structure of the adjectives attributed to students. They found women were more likely to be described as "compassionate," and "enthusiastic," whereas male students were more likely to be described as "quick learners."[1] Within the field of Emergency Medicine (EM), researchers examined qualitative feedback for residents and found gendered differences in the written feedback male and female residents received. Specifically, men received concordant feedback when their performance struggled, but women tended to receive discordant feedback about what aspects of their performance needed to improve.[5] For example, while one attending evaluator praised a female resident's responsiveness to feedback, another evaluator commented that the resident needed to be more open to feedback. In addition, this analysis suggested that attending evaluators found the ideal EM resident to possess many stereotypically masculine traits.[5] These examples at the medical student and resident training levels suggest that gender bias exists within performance evaluations.

These issues of gender bias are particularly relevant in the field of surgery, where women remain the minority of resident trainees despite women having achieved gender parity in medical schools.[6,7] Data show that gender bias is a widespread problem in surgical residency training. A 2015 survey found that 88% of female residents perceived gender-based discrimination during their surgical residency.[8]

Regular feedback has been shown to be very important to surgical trainees, and feedback in the form of written performance evaluations is a routine part of surgical residency training.[9]

* Corresponding author. Washington University School of Medicine, 4901 S Euclid Ave Suite 920, St. Louis, MO, 63108, United States.
E-mail address: arghavan@alumni.stanford.edu (A. Salles).
[1] These authors contributed equally to this work.
[2] Present Address: Washington University School of Medicine, 4901 S Euclid Ave Suite 920, St. Louis, MO 63108, (P) 314-747-8877.

https://doi.org/10.1016/j.amjsurg.2018.09.029
0002-9610/

However, no group has analyzed the language of written surgical resident evaluations for gender bias. In this study, our goal was to discover the ways people may write differently about men and women surgical trainees through the use of a grounded theory framework.

Methods

After obtaining permission from the Institutional Review Board at Stanford University all surgical residents (N = 166) at a tertiary care center were invited to participate in a voluntary longitudinal study of resident performance and well-being from 2010 to 2011. Within the context of this study, we obtained performance evaluations of residents in nine surgical specialties: cardiothoracic surgery, general surgery, neurosurgery, ophthalmology, orthopedic surgery, otolaryngology, plastic surgery, urology, and vascular surgery. In addition to quantitative scores, all evaluations contained a section asking reviewers to comment in free-text about the resident's performance during the rotation. The stem for this section varied slightly amongst departments, but generally the stem was worded as, "Overall assessment of the resident" or "Overall comments about the resident." While some evaluations seem to be based on one individual's input, others appear to be composite feedback from a group of individuals. All evaluations were de-identified, and gender pronouns were removed to blind researchers to the resident gender. We used a grounded theory approach as described by Glaser & Strauss to analyze the narrative performance evaluation data with inductive thematic analysis to discover themes that emerged from the data.[10–12] We also performed a content analysis based on previously determined words and categories.[13,14]

A grounded theory approach was used for coding of the content in written performance evaluations as well as generation of a conceptual model.[10] Grounded theory offers a rigorous approach to qualitative data through the identification and application of a coding scheme for the raw data, organization of codes into important themes, and analysis of these themes; thus, the themes arise from the data, rather than any prior assumptions or hypotheses about the data.[10]

Inductive thematic analysis

We analyzed the written comments using a three-step process: open coding (bits of meaning that generate codes), followed by axial coding (defining themes), and concluding with selective coding (overarching model of the themes).[15,16] In open coding, two researchers (KS and JM) independently read a randomly selected subset of all of the comments and individually coded each idea within each comment. We coded enough evaluations to reach thematic saturation which occurred after 50 evaluations were analyzed.[10] We met to reconcile differences in coding between the two raters, and a codebook was developed to analyze the evaluations. Two other evaluators (KG and ML) were trained on the codebook and analyzed the remaining data. They did not identify any additional codes upon analysis of the rest of the dataset. Both KG and ML independently coded every evaluation in the dataset.

In addition to these codes, each idea was assigned a valence, using a −2 to +2 scale. An example of a very positive comment (+2) is "*outstanding* attention to details" while a very negative comment (−2) is "*Frequently* missed *significant* details". For each code for each resident, the sum of the valence within that code was recorded as the magnitude of the evaluation with regard to that code. The sum across all codes for an individual resident's evaluation was used as a measure of the overall tone toward that resident. Although valence is not routinely used in grounded theory, we evaluated valence in order to assess trends in overall tone. Coding disagreements were resolved through discussion.

Following open coding, axial coding was conducted by classifying the codes into meaningful themes.[15] The investigators communicated regularly during all stages of coding and classification, which allowed the investigators to challenge ideas and raise insights. This process culminated in a conceptual model describing how people write differently about men and women surgical trainees.

Content analysis

Because previous research has identified differential use of specific words to describe men and women, we examined the incidence of a number of specific terms, which we compiled from existing lists to perform a content analysis.[13,14] We assessed the use of communal, grindstone, standout, and ability words which were previously reported in the literature.[17–22] The specific words we examined in each of these four categories are listed in Table 1. Communal terms are generally associated with community building and interpersonal dynamics. Grindstone terms described the resident's effort put forth. Standout terms are more typically associated with an individual's noteworthy performance. Ability terms refer to terms that are used to describe intrinsic characteristics of the resident.

We created a novel measure called relative use, which is similar to relative risk, to assess the differential use of words within each specific category (communal, grindstone, standout, and ability terms) by gender:

$$Relative\ Use = \frac{\left(\frac{number\ of\ category\ words\ used\ for\ men}{number\ of\ words\ in\ evaluations\ of\ men}\right)}{\left(\frac{number\ of\ category\ words\ used\ for\ women}{number\ of\ words\ in\ evaluations\ of\ women}\right)}$$

For each gender, we calculated the number of times each word within each category was used divided by the total number of words in evaluations for that gender. We then summed up the frequencies for all words within each category. A relative use greater than one is associated with more frequent use of a word in evaluations of men, while a relative use less than one means more frequent use of a word in evaluations of women. Although this not a validated measure, it does have face validity and allowed us to take into account the fact that there were more evaluations written about men and that these words were sometimes used multiple times within an evaluation. Fisher's exact tests were performed to compare the relative use of each word between men and women using SAS (Version 9.4). Alpha was set at 0.05 for all tests.

Results

As shown in Table 2, we obtained 539 resident performance evaluations (197 women, 347 men) from 143 unique residents (representing 86% of all surgical residents) within 9 surgical subspecialties at one institution. Residents represented the full range of years of training and included 92 men and 51 women.

Inductive thematic analysis

As shown in Fig. 1, the thematic analysis yielded 32 unique codes, which we categorized into seven themes plus two stand-alone codes. The two individual codes were Overall Performance and Reference to Future, and the themes were Job Domains, Professional Competency, Disposition and Humanism, Work Efficacy, Growth, Insight, and Team Dynamic. Representative comments for the seven themes and two stand-alone codes can be found in

Table 1
Word lists from the literature used for content analysis.

Communal	Grindstone	Ability	Standout
Caring[19]	Conscientious[22]	Intelligent[22]	Exceptional[22]
Kind[19]	Diligent[19]	Bright[22]	Best[18]
Empathy[19]	Meticulous[22]	Talent/Talented[22]	Outstanding[22]
Compassionate[19]	Organize/organized/organization[22]	Brilliant[22]	Superb[19]
Communicate/communicated[22]	Tireless/tirelessly[19]	Competent[18]	Excellent[22]
Rapport[22]	Solid[19]	Smart[22]	Phenomenal[19]
		Gifted[22]	Star/superstar[19]
			Terrific[22]
			Fabulous[22]
			Leader[22]
			Scholar/Scholarly[22]

Table 3. Inter-rater agreement was 85% for codes and 87.5% for valence. Inter-rater reliability (Cohen's kappa) was 0.84.

Overall performance

Men tended to receive more positive comments than women about their overall performance, and men were more likely to be described in a superlative manner. For example, one evaluation for a man read,

> "*Terrific* resident. Great job! X is *a star* … He is performing quite well this year for his level of training." (PGY 1)

This particular evaluation is a composite from multiple evaluators, and it is quite clear they all have a very high opinion of him. Evaluations of women consistently had fewer words and phrases suggesting outstanding or exemplary performance:

> "X is a good resident. She has a good knowledge base, works hard and thinks about what she is doing." (PGY 3)

These two comments are representative of a trend in which men received comments on their overall performance that were more positive than were comments about women, e.g. "terrific" as opposed to "good." Results from our content analysis of standout words (see below) complement this finding.

Reference to future

Similarly, when making a reference to the future potential of residents, faculty typically described the men more positively than the women. Differences in the ideas coded as Reference to Future were consistent both early and later in training. A representative comment about a woman resident is:

> "*Seemed to be* well grounded *for an intern* in first rotation. Has good attitude and potential." (PGY 1)

The passive tone, referring to how she *seemed*, makes this evaluation less strong than an evaluation of the same traits in a male resident:

> "Excellent intern for first month rotation. Will be a leader in our program." (PGY 1)

He *is* excellent in his first rotation, and there is more confidence in his future as a leader than there is in her potential. The use of superlatives tended to be more frequent in evaluations of male residents across the board, particularly with respect to how their future was perceived. A PGY 5 male resident was described as

> "a superb resident - one of the best I have worked with over the last 23 years. *I predict a great future for him* as an academic neurosurgeon/scientist and leader nationally."

Table 2
Distribution of resident participants by subspecialty and by PGY.

	Men	Women
	No. of residents (No. of evals)	No. of residents (No. of evals)
Surgical Subspecialty		
Cardiothoracic Surgery	4 (12)	1 (2)
General Surgery	18 (62)	18 (73)
Neurosurgery	14 (25)	2 (4)
Otolaryngology Head & Neck Surgery	12 (116)	8 (58)
Ophthalmology	6 (9)	5 (9)
Orthopedic Surgery	17 (24)	6 (9)
Plastic Surgery	12 (77)	5 (18)
Urology	8 (19)	5 (14)
Vascular Surgery	1 (3)	1 (5)
Total	**92 (347)**	**51 (192)**
Post-Graduate Year		
1	20 (73)	10 (33)
2	20 (96)	10 (40)
3	19 (50)	14 (60)
4	17 (63)	9 (31)
5	13 (57)	8 (28)
6	3 (8)	0 (0)
Total	**92 (347)**	**51 (192)**

Fig. 1. Codes grouped into themes from inductive thematic analysis. The numbers represent the number of times a code was used in the dataset.

A PGY3 male resident received similar feedback:

"He is a superb resident. He has *excellent potential* to become a leader in academic neurosurgery."

However, his female colleague (PGY3) was

"a great resident. *She has potential to succeed* as an academic neurosurgeon."

Her evaluator seems less certain of her future success than the evaluator of the male residents; she has potential to succeed, but her evaluator is not considering her success to be a certainty, nor is the evaluator categorizing her as a future leader.

Job domains

Comments within the Job Domains theme tended to focus more on exemplary and commendable traits in evaluations of male residents (as compared to evaluations of female residents). For example, with regard to motivation, a man in his 2nd year of ophthalmology residency, who was "below average for [his] level of training but aware of [his] own limits," was held up as a role model, being "highly motivated to do extra reading on his own time, a quality that should be emulated by other residents." A female trainee was described as "an outstanding surgeon … ahead of the curve as a 2nd year resident," and her "orthopaedic knowledge base is excellent and she is a motivated and enthusiastic learner." The subpar male trainee was a role model for his commitment to learning, but no such comment was made about the talented and skilled woman. While this could be reflective of the individual evaluators' tendencies rather than the residents, this was a notable pattern throughout the evaluations we examined.

Professional competency

The comments in the theme for Professional Competency demonstrated some subtler differences. A female urology resident was "very professional and *pleasant to deal with*," (PGY4), whereas her male colleague in urology was "thoughtful and conducts himself professionally … *a real gentleman*" (PGY3). Superficially, these comments are similar: the residents are both commended for their professionalism and demeanor. Upon a closer examination, there is

Table 3
Themes identified with representative comments.

Theme	Representative Comment
Job Domains	"Excellent bedside manners, clinical acumen and surgical skills."
Professional Competency	"X was always punctual, well prepared for the surgical cases"
Team Dynamic	"Would find creative ways for residents on the team to all obtain productive experiences. Would stay till the TEAM's work was done, not just when X's case was done."
Work Efficacy	"Works hard and sees all tasks to completion. Outstanding in initiating activity and taking care of problems."
Disposition and Humanism	"Very personable and establishes rapport with patients easily, compassionate in his/her care for ophthalmic patients"
Growth	"Accepts feedback well and works to assimilate it to patient care. X is mature and reflective."
Insight	"appropriately self-aware and self-critical"
Overall Performance	"Outstanding resident. One of the top two this year."
Reference to Future	"X will do fine in his/her future practice"

a negative connotation associated with being someone "to deal with" rather than a person with a desirable quality (being a "gentleman"). This subtle language is an example of a common occurrence in the data, undermining female residents' performance. For example, another female intern's evaluation began by noting that, "she has been a *delightful surprise*." The rest of her evaluation was very positive, noting her enthusiasm, efficiency, operative skills, dedication, and attitude, but it was tempered by the inadvertent jab at how unexpected her strong performance was.

Disposition and humanism

In addition, we noticed certain phrases that were only written about women; these stood out among the 197 evaluations of female residents, particularly in the theme of Disposition and Humanism. These words and phrases were exclusively seen in the evaluations of women:

"always smiling" (PGY3)

"an absolute gem" (PGY4)

"interacts with everyone in a pleasant demeanor and *never seems to get upset or angry*" (PGY4)

"Solid, consistent work *with no fuss*." (PGY1)

Comments like these may undermine the professionalism and job-related skills of these female surgeons. Sentiments regarding demeanor, attitude, professionalism, and teamwork were communicated in more appropriate, professional terms in evaluations of men. The following are examples of this:

"Hard working, pleasant resident." (PGY2)

"Calm and determined demeanor in difficult situations" (PGY5)

These phrases depict the residents' strengths and performance on their rotation without implying that these attributes were unexpected.

Overall tone and remaining themes

The final notable difference in evaluations for men and women was in the overall tone: evaluations of men tended to be more positive than the evaluations of women. On average, the tone value for men was 26.9 (SE = 2.55) while the tone value for women was 16.15 (SE = 2.068, $t = 3.29$, $p = 0.001$.) There was no statistically significant gender-based difference in the number of words used per evaluation: men's evaluations averaged 175.3 words in length, and women's evaluations averaged 183.3 words in length, ($t = -0.26$, $p = 0.795$).

No differences were obvious in the remaining themes: Work Efficacy, Growth, Insight, and Team Dynamic, with an exception for the leadership code, which is addressed in the content analysis below.

Upon reflection of our data and the themes that emerged from selective coding, a conceptual model emerged, which describes how people write differently about men and women surgical trainees. Fig. 2 depicts our model, which divides the content in resident evaluations into gendered and non-gendered constructs.

Content analysis

To further investigate the evaluations, we examined how word choice differed in evaluations of men and women. Overall, there were 17,012 words written about the 92 men and 9374 words written about the 51 women. As shown in Table 4, we tabulated the total number of occurrences of each key communal, grindstone, ability, and standout word by gender. While there were no statistically significant differences in relative use for communal, grindstone, and ability words, men were more likely than women to be described using standout words (relative use = 1.40, $p < 0.01$). Within this category, the specific standout word "leader" was used significantly more often in men's evaluations (relative use = 3.1, $p = 0.02$) than in women's evaluations.

Discussion

This work is the first to qualitatively study surgical resident written performance evaluations. In the evaluations of men and women surgical residents, several thematic differences emerged between genders, which ultimately culminated in a conceptual model describing how people write differently about men and women surgical trainees. Specifically, men received comments that were more positive regarding their overall performance and future potential, while women's evaluations were more likely to contain certain modulating phrases that were never written in evaluations of men. These phrases tended to introduce uncertainty or undermine the women's performance. Additionally, the types of words used to describe men and women residents also differed, with men being more frequently described using standout words including the word leader.

Our conceptual model describes how content in resident evaluations can be divided into 2 categories: themes with gendered use, and themes without gendered use. Upon further analysis of these 2 categories, we find that the themes with gendered use tend to judge the residents' current and future ability. Conversely, themes without gendered use tended to do with personality and motivation. Our finding that ability is a gendered construct is consistent with Trix et al., who found that men were more likely to be described using ability words in letters of recommendation.[21]

Our findings suggest that the ways men and women are written about in surgical residency evaluations differ in meaningful ways. In a prior study, our group examined the quantitative aspects of these performance evaluations and found no significant difference between genders.[23] Despite no quantitative difference in evaluations, we found a qualitative difference in multiple themes in the comments of the same evaluations. Implicit bias may explain why men and women are described differently in narrative feedback. An alternative explanation could be that men simply had higher job performance. However, given the lack of a quantitative difference between these groups, the latter explanation seems less likely. In addition, other studies have shown that women physicians have equal, if not superior, outcomes compared to men.[24,25]

Our content analysis was consistent with the literature regarding standout words. Prior work in this area has shown that men were more likely than women to be described using standout words in letters of recommendation for chemistry/biochemistry jobs.[22] Just as in our data, a gender difference in the description of applicants existed despite similar levels of objective qualifications. Our finding that the length of men's and women's evaluations did not differ contrasts to the findings of Trix et al., who showed that letters of recommendation written for men were significantly longer than those written for women.[21] Letters of recommendation serve a different purpose than evaluations, and that may explain this difference.

Our data are consistent with a recent qualitative analysis of R01 grant renewal proposals. Researchers found that male investigators were more likely to be described as leaders and pioneers while female investigators were commended for expertise and the quality of their environments.[26] Men's successes were attributed to individual talents, while women's successes tended to be attributed

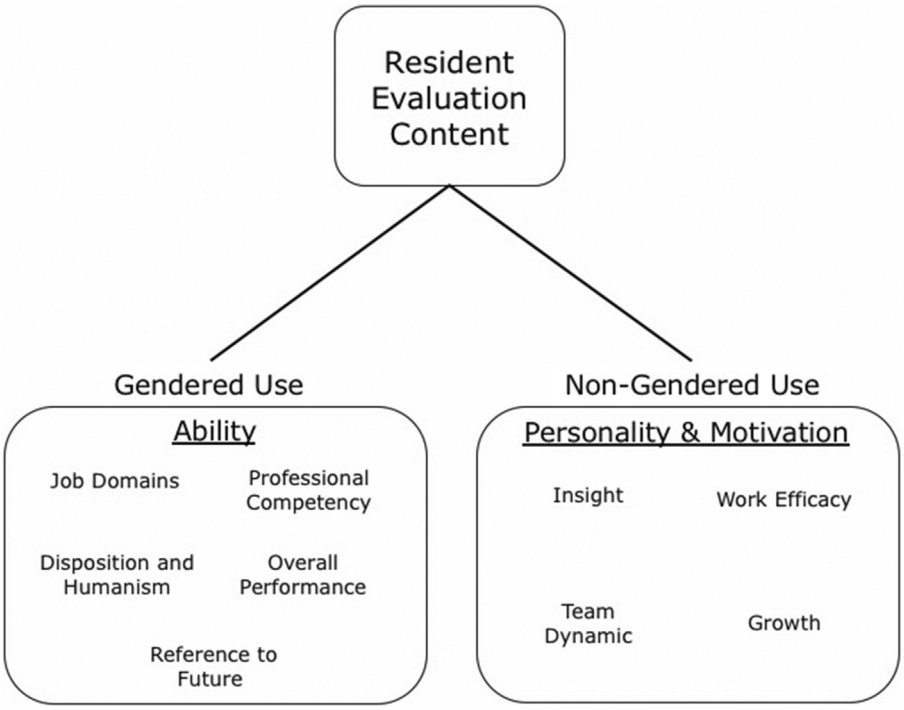

Fig. 2. Conceptual model of how people write differently about men and women surgical trainees.

Table 4
Content analysis.

Word Category	Word	Unadjusted Word Counts		Adjusted Relative Use[+] *(men/women)*	p-value
		Men	*Women*		
Communal	**Total**	**49**	**27**	**1.00**	**1.00**
Grindstone	**Total**	**71**	**34**	**1.15**	**0.54**
Ability	**Total**	**54**	**21**	**1.42**	**0.19**
Standout	Excellent	169	72	1.29	0.07
	Outstanding	69	25	1.52	0.08
	Leader	28	5	3.09	0.02
	Best	15	11	0.75	0.54
	Exceptional	13	6	1.19	0.81
	Superb	13	4	1.79	0.45
	Star/Superstar	7	2	1.93	0.51
	Terrific	8	1	4.41	0.17
	Scholar/Scholarly	5	0	*	*
	Phenomenal	1	1	0.55	1.00
	Fabulous	0	2	*	*
	Total	**328**	**129**	**1.40**	**<0.01**

[+] Adjusted use was the ratio of two proportions: $Relative\ Use = \frac{\left(\frac{number\ of\ category\ words\ used\ for\ men}{number\ of\ words\ in\ evaluations\ of\ men}\right)}{\left(\frac{number\ of\ category\ words\ used\ for\ women}{number\ of\ words\ in\ evaluations\ of\ women}\right)}$.

* Could not be calculated because of zero-use for one gender.

to working hard and having support around them. This was seen to an extent in our data, as women were often described as having a potential to succeed while men were simply expected to succeed. The R01 study also noted that applications from men received higher scores overall, which differs from what was found in our previous quantitative study but is consistent with our finding that the tone of evaluations for men was consistently more positive than that for women.[26]

Data presented in the current study are similar to what has been found in other fields as well. A 2014 study conducted by Fortune magazine found that, compared to feedback for men, feedback for women was far more likely to focus on their personalities: 71 of 97 critical reviews of women remarked on personality, while only 2 of 83 critical reviews of men included similar comments.[27] In addition, while men were given explicit suggestions, like "hone your strategies for guiding your team and developing their skills," women were given more vague feedback, such as "sometimes you need to step back to let others shine."[27] Similarly, in a report on women in the workplace, women were found to be 22% less likely than men to receive feedback that could be used to improve their performance.[28] Finally, venture capitalists speak differently about men and women behind closed doors, referring to men as "young

and promising", while calling women of similar backgrounds "young, but inexperienced," which likely contributes to women's lower rate of success as entrepreneurs.[29]

There are limitations to our qualitative study. Our data was gathered from a single academic year at one institution, with slightly more evaluations written for residents earlier in their training. Additionally, we only have one source of data. However, multiple researchers analyzed the data, allowing for multiple perspectives on the data. Although it is possible that we did not reach thematic saturation in our dataset, the fact that no additional themes were identified beyond the first 50 evaluations suggests we may have. In our study, several variables were unknown, including the gender of the evaluator and the race or ethnicity of the evaluator and the resident. Various studies in this field have demonstrated no effect with respect to evaluator gender.[23,30–33]

A further limitation in our study was the lack of context for the comments in the evaluations we read. We were therefore unable to tell whether individual comments about inadequate performance were justified by specific incidents. That said, our sample size was sufficiently large to enable us to describe tendencies that persist beyond one or two individual instances.[23] Additionally, research in emergency medicine found that many of the characteristics valued in EM residency were stereotypically associated with men.[5] This could certainly have an impact on how men and women in surgical residency are perceived.

Our findings suggest that there is room for improvement with regard to implicit bias and its impact on the training of surgical residents. Furthermore, bias likely occurs at various levels of training and promotion. Future research should study performance evaluations at all levels, including Medical Student Performance Evaluations (MSPEs) and evaluations of faculty. Other forms of bias, including racial, ethnic, sexual orientation, weight, and religious bias are also likely to be present, so future research should explicitly study these biases in performance evaluations as well. We would encourage residency training programs to thoughtfully review their programs' performance evaluations for patterns of bias. Programs may consider implementing bias training, which has been shown to be an effective method of improving faculty awareness of bias issues and increasing hiring diversity in STEMM (science, technology, engineering, math, and medicine) departments.[34,35] While some data on bias trainings are promising, other data are mixed.[36–38] Therefore, future studies should investigate the most effective way to mediate multiple forms of bias and their manifestations in performance evaluations.

Funding sources

This research was partially funded by the National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH, through Grant 5 KL2 RR025743.

Table of contents summary

In the qualitative analysis of men and women surgical residents' performance evaluations, there were notable differences between genders. These differences emerged in the themes of overall performance, references to the future, professional competency, job domains, and disposition and humanism. The overall tone of evaluations was also more positive for men than for women. Finally, evaluations of men included more standout words. The assessment of surgical residents may be implicitly gender biased.

Acknowledgements

The authors would like to think Laurel Milam for her assistance with the statistical analyses for this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amjsurg.2018.09.029.

References

1. Axelson RD, Solow CM, Ferguson KJ, Cohen MB. Assessing implicit gender bias in medical student performance evaluations. Eval Health Prof. 33(3):365-385. doi:10.1177/0163278710375097.
2. Thackeray EW, Halvorsen AJ, Ficalora RD, Engstler GJ, McDonald FS, Oxentenko AS. The effects of gender and age on evaluation of trainees and faculty in gastroenterology. *Am J Gastroenterol*. 2012;107(11):1610–1614. https://doi.org/10.1038/ajg.2012.139.
3. Rand VE, Hudes ES, Browner WS, Wachter RM, Avins AL. Effect of evaluator and resident gender on the American Board of Internal Medicine evaluation scores. *J Gen Intern Med*. 1998;13(10):670–674. https://doi.org/10.1046/j.1525-1497.1998.00202.x.
4. Choo EK. Damned if you do, damned if you don't: bias in evaluations of female resident physicians. *J Grad Med Educ*. 2017;9(5):586–587. https://doi.org/10.4300/JGME-D-17-00557.1.
5. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender differences in attending physicians' feedback to residents: a qualitative analysis. *J Grad Med Educ*. 2017;9(5):577–585. https://doi.org/10.4300/JGME-D-17-00126.1.
6. AAMC. *ACGME Residents and Fellows by Sex and Specialty*; 2015. https://www.aamc.org/data/workforce/reports/458766/2-2-chart.html. Accessed June 15, 2018.
7. AAMC. *More Women than Men Enrolled in U.S. Medical Schools in 2017*; 2017. https://news.aamc.org/press-releases/article/applicant-enrollment-2017/. Accessed June 15, 2018.
8. Bruce AN, Battista A, Plankey MW, Johnson LB, Marshall MB. Perceptions of gender-based discrimination during surgical training and practice. *Med Educ Online*. 2015;20:25923. http://www.ncbi.nlm.nih.gov/pubmed/25652117. Accessed June 15, 2018.
9. Bello RJ, Sarmiento S, Meyer ML, et al. Understanding surgical resident and fellow perspectives on their operative performance feedback needs: A Qualitative Study. doi:10.1016/j.jsurg.2018.04.002.
10. Glaser BG, Strauss AL. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine de Gruyter; 1967. http://www.worldcat.org/title/discovery-of-grounded-theory-strategies-for-qualitative-research/oclc/253912. Accessed September 1, 2018.
11. Miles MB, Huberman AM. *Qualitative Data Analysis: A Sourcebook of New Methods*. Newbury Park, CA: Sage Publications, Inc.; 1984.
12. Nowell LS, Norris JMM, White DEE, Moules NJJ. Thematic analysis. *Int J Qual Methods*. 2017;16(1). https://doi.org/10.1177/1609406917733847, 160940691773384.
13. Woodrum E. "Mainstreaming" content analysis in social science: methodological advantages, obstacles, and solutions. *Soc Sci Res*. 1984;19(1):1–19. https://www-sciencedirect-com.beckerproxy.wustl.edu/science/article/pii/0049089X84900012. Accessed September 1, 2018.
14. Holsti OR. Content analysis. In: Lindzey G, Aronson E, eds. *The Handbook of Social Psychology*. second ed. vol. 2. Reading, MA: Addison-Wesley Publishing Co.; 1968.
15. Corbin J, Strauss A. Grounded theory research: procedures, canons, and evaluative criteria. *Qual Sociol*. 1990;13:3–21. http://med-fom-familymed-research.sites.olt.ubc.ca/files/2012/03/W10-Corbin-and-Strauss-grounded-theory.pdf. Accessed September 1, 2018.
16. Cochran A, Elder WBB. A model of disruptive surgeon behavior in the perioperative environment. *J Am Coll Surg*. 2014;219(3):390–398. https://doi.org/10.1016/j.jamcollsurg.2014.05.011.
17. Madera JM, Hebl MR, Martin RC. Gender and letters of recommendation for academia: agentic and communal differences. *J Appl Psychol*. 2009;94(6):1591–1599. https://doi.org/10.1037/a0016539.
18. Ross DA, Boatright D, Nunez-Smith M, Jordan A, Chekroud A, Moore EZ. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. Gold JA, ed. *PloS One*. 2017;12(8). https://doi.org/10.1371/journal.pone.0181659. e0181659.
19. Messner AH, Shimahara E. Letters of recommendation to an otolaryngology/head and neck surgery residency program: their function and the role of gender. *Laryngoscope*. 2008;118(8):1335–1344. https://doi.org/10.1097/MLG.0b013e318175337e.
20. Isaac C, Chertoff J, Lee B, Carnes M. Do students' and authors' genders affect evaluations? A linguistic analysis of medical student performance evaluations. *Acad Med*. 2011;86(1):59–66. https://doi.org/10.1097/ACM.0b013e318200561d.
21. Trix F, Psenka C. Exploring the color of glass: letters of recommendation for female and male medical faculty. *Discourse Soc*. 2003;14(2):191–220. https://doi.org/10.1177/0957926503014002277.
22. Schmader T, Whitehead J, Wysocki VH. A linguistic comparison of letters of

recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*. 2007;57(7-8):509–514. https://doi.org/10.1007/s11199-007-9291-4.
23. Salles A, Mueller CM, Cohen GL. A values affirmation intervention to improve female residents' surgical performance. *J Grad Med Educ*. 2016;8(3):378–383. https://doi.org/10.4300/JGME-D-15-00214.1.
24. Wallis CJ, Ravi B, Coburn N, Nam RK, Detsky AS, Satkunasivam R. Comparison of postoperative outcomes among patients treated by male and female surgeons: a population based matched cohort study. *BMJ*. October 2017:j4366. https://doi.org/10.1136/bmj.j4366.
25. Tsugawa Y, Jena AB, Figueroa JF, Orav EJ, Blumenthal DM, Jha AK. Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians. *JAMA Intern Med*. 2017;177(2):206–213. https://doi.org/10.1001/jamainternmed.2016.7875.
26. Magua W, Zhu X, Bhattacharya A, et al. Are female applicants disadvantaged in national institutes of health peer review? Combining algorithmic text mining and qualitative methods to detect evaluative differences in R01 reviewers' critiques. *J Wom Health*. 2017;26(5):560–570. https://doi.org/10.1089/jwh.2016.6021.
27. Snyder K. Performance review gender bias: high-achieving women are "abrasive.". *Fortune*; 2014. http://fortune.com/2014/08/26/performance-review-gender-bias/. Accessed April 10, 2018.
28. LeanIn, McKinsey & Company. Women in the workplace. *Leanin.org*. 2015. https://doi.org/10.1111/0045-3609.00002.
29. Malmström M, Johansson J, Wincent J. Gender stereotypes and venture support decisions: how governmental venture capitalists socially construct entrepreneurs' potential. *Enterpren Theor Pract*. 2017;41(5):833–860. https://doi.org/10.1111/etap.12275.
30. Devine PG. Stereotypes and prejudice: their automatic and controlled components. *J Pers Soc Psychol*. 1989;56(1):5–18. https://doi.org/10.1037/0022-3514.56.1.5.
31. Nosek BA, Smyth FL, Hansen JJ, et al. Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur Rev Soc Psychol*. 2007;18(1):36–88. https://doi.org/10.1080/10463280701489053.
32. Blanch DC, Hall JA, Roter DL, Frankel RM. Medical student gender and issues of confidence. *Patient Educ Counsel*. 2008;72(3):374–381. https://doi.org/10.1016/j.pec.2008.05.021.
33. Dayal A, O'Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. *JAMA Intern Med*. 2017;177(5):651–657. https://doi.org/10.1001/jamainternmed.2016.9616.
34. Carnes M, Devine PG, Baier Manwell L, et al. The effect of an intervention to break the gender bias habit for faculty at one institution. *Acad Med*. 2015;90(2):221–230. https://doi.org/10.1097/ACM.0000000000000552.
35. Devine PG, Forscher PS, Cox WTL, Kaatz A, Sheridan J, Carnes M. A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *J Exp Soc Psychol*. 2017;73:211–215. https://doi.org/10.1016/j.jesp.2017.07.002.
36. Dobbin F, Schrage D, Kalev A. Rage agains the iron cage: the varied effects of bureaucratic personnel reforms on diversity. *Am Sociol Rev*. 2015;80(5):1014–1044. https://doi.org/10.1177/0003122415596416.
37. Duguid MM, Thomas-Hunt MC. Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. *J Appl Psychol*. 2015;100(2):343–359. https://doi.org/10.1037/a0037908.
38. Atewologun D, Cornish T, Tresh F. Unconscious bias training: an assessment of the evidence for effectiveness equality and human rights commission research report series.. www.equalityhumanrights.com; 2018. Accessed August 30, 2018.